

# Insights from Classifying Visual Concepts with Multiple Kernel Learning

Alexander Binder<sup>\*†</sup>, Shinichi Nakajima<sup>‡</sup>, Marius Kloft<sup>§</sup>, Christina Müller, Wojciech Samek,  
Ulf Brefeld<sup>¶</sup>, Klaus-Robert Müller<sup>||</sup>, and Motoaki Kawanabe<sup>\*\*</sup>

December 19, 2011

## Abstract

Combining information from various image features has become a standard technique in concept recognition tasks. However, the optimal way of fusing the resulting kernel functions is usually unknown in practical applications. Multiple kernel learning (MKL) techniques allow to determine an optimal linear combination of such similarity matrices. Classical approaches to MKL promote sparse mixtures. Unfortunately, so-called 1-norm MKL variants are often observed to be outperformed by an unweighted sum kernel. The contribution of this paper is twofold: We apply a recently developed non-sparse MKL variant to state-of-the-art concept recognition tasks within computer vision. We provide insights on benefits and limits of non-sparse MKL and compare it against its direct competitors, the sum kernel SVM and the sparse MKL. We report empirical results for the PASCAL VOC 2009 Classification and ImageCLEF2010 Photo Annotation challenge data sets. About to be submitted to PLoS ONE.

## 1 Introduction

A common strategy in visual object recognition tasks is to combine different image representations to capture relevant traits of an image. Prominent representations are for instance built from color, texture, and shape information and used to accurately locate and classify the objects of interest. The importance of such image features changes across the tasks. For example, color information increases the detection rates of stop signs in images substantially but it is almost useless for finding cars. This is because stop sign are usually red in most countries but cars in principle can have any color. As additional but nonessential features not only slow down the computation time but may even harm predictive performance, it is necessary to combine only relevant features for state-of-the-art object recognition systems.

We will approach visual object classification from a machine learning perspective. In the last decades, support vector machines (SVM) [1, 2, 3] have been successfully applied to many practical problems in various fields including computer vision [4]. Support vector machines exploit similarities of the data, arising from some (possibly nonlinear) measure. The matrix of pairwise similarities, also known as kernel matrix, allows to abstract the data from the learning algorithm [5, 6].

That is, given a task at hand, the practitioner needs to find an appropriate similarity measure and to plug the resulting kernel into an appropriate learning algorithm. But what if this similarity measure is difficult to find? We note that [7] and [8] were the first to exploit prior and domain knowledge for the kernel construction.

In object recognition, translating information from var-

<sup>\*</sup>corresponding author, alexander.binder@tu-berlin.de

<sup>†</sup>A. Binder and W. Samek are with Technische Universität Berlin and Fraunhofer FIRST, Berlin, Germany

<sup>‡</sup>S. Nakajima is with Optical Research Laboratory, Nikon Corporation Tokyo

<sup>§</sup>M. Kloft and C. Müller are with Technische Universität Berlin, Germany

<sup>¶</sup>U. Brefeld is with the Universität Bonn, Germany

<sup>||</sup>K.-R. Müller is with Technische Universität Berlin, Germany, and the Institute of Pure and Applied Mathematics at UCLA, Los Angeles, USA

<sup>\*\*</sup>M. Kawanabe is with ATR Research, Kyoto, Japan

ious image descriptors into several kernels has now become a standard technique. Consequently, the choice of finding the right kernel changes to finding an appropriate way of fusing the kernel information; however, finding the right combination for a particular application is so far often a matter of a judicious choice (or trial and error).

In the absence of principled approaches, practitioners frequently resort to heuristics such as uniform mixtures of normalized kernels [9, 10] that have proven to work well. Nevertheless, this may lead to sub-optimal kernel mixtures.

An alternative approach is multiple kernel learning (MKL) that has been applied to object classification tasks involving various image descriptors [11, 12]. Multiple kernel learning [13, 14, 15, 16] generalizes the support vector machine framework and aims at learning the optimal kernel mixture and the model parameters of the SVM simultaneously. To obtain a well-defined optimization problem, many MKL approaches promote sparse mixtures by incorporating a 1-norm constraint on the mixing coefficients. Compared to heuristic approaches, MKL has the appealing property of learning a kernel combination (wrt. the  $\ell_1$ -norm constraint) and converges quickly as it can be wrapped around a regular support vector machine [15]. However, some evidence shows that sparse kernel mixtures are often outperformed by an unweighted-sum kernel [17]. As a remedy, [18, 19] propose  $\ell_2$ -norm regularized MKL variants, which promote non-sparse kernel mixtures and subsequently have been extended to  $\ell_p$ -norms [20, 21].

Multiple Kernel approaches have been applied to various computer vision problems outside our scope such multi-class problems [22] which require mutually exclusive labels and object detection [23, 24] in the sense of finding object regions in an image. The latter reaches its limits when image concepts cannot be represented by an object region anymore such as the *Outdoor*, *Overall Quality* or *Boring* concepts in the ImageCLEF2010 dataset which we will use.

In this contribution, we study the benefits of sparse and non-sparse MKL in object recognition tasks. We report on empirical results on image data sets from the PASCAL visual object classes (VOC) 2009 [25] and ImageCLEF2010 PhotoAnnotation [26] challenges, showing that non-sparse MKL significantly outperforms the uniform mixture and  $\ell_1$ -norm MKL. Furthermore we discuss

the reasons for performance gains and performance limitations obtained by MKL based on additional experiments using real world and synthetic data.

The family of MKL algorithms is not restricted to SVM-based ones. Another competitor, for example, is Multiple Kernel Learning based on Kernel Discriminant Analysis (KDA) [27, 28]. The difference between MKL-SVM and MKL-KDA lies in the underlying single kernel optimization criterion while the regularization over kernel weights is the same.

Outside the MKL family, however, within our problem scope of image classification and ranking lies, for example, [29] which uses a logistic regression as base criterion and results in a number of optimization parameters equal to the number of samples times the number of input features. Since the approach in [29] uses a priori much more optimization variables, it poses a more challenging and potentially more time consuming optimization problem which limits the number of applicable features and can be evaluated for our medium scaled datasets in detail in the future.

Alternatives use more general combinations of kernels such as products with kernel widths as weighting parameters [30, 31]. As [31] point out the corresponding optimization problems are no longer convex. Consequently they may find suboptimal solutions and it is more difficult to assess using such methods how much gain can be achieved via learning of kernel weights.

This paper is organized as follows. In Section 2, we briefly review the machine learning techniques used here; The following section3 we present our experimental results on the VOC2009 and ImageCLEF2010 datasets; in Section 4 we discuss promoting and limiting factors of MKL and the sum-kernel SVM in three learning scenarios.

## 2 Methods

This section briefly introduces multiple kernel learning (MKL), and kernel target alignment. For more details we refer to the supplement and the cited works in it.

## 2.1 Multiple Kernel Learning

Given a finite number of different kernels each of which implies the existence of a feature mapping  $\psi_j : \mathcal{X} \rightarrow \mathcal{H}_j$  onto a hilbert space

$$k_j(\mathbf{x}, \bar{\mathbf{x}}) = \langle \psi_j(\mathbf{x}), \psi_j(\bar{\mathbf{x}}) \rangle_{\mathcal{H}_j}$$

the goal of multiple kernel learning is to learn SVM parameters  $(\alpha, b)$  and linear kernel weights  $K = \sum_l \beta_l k_l$  simultaneously.

This can be cast as the following optimization problem which extends support vector machines [2, 6]

$$\begin{aligned} \min_{\beta, v, b, \xi} \quad & \frac{1}{2} \sum_{j=1}^m \frac{v'_j v_j}{\beta_j} + C \|\xi\|_1 \\ \text{s.t.} \quad & \forall i : y_i \left( \sum_{j=1}^m v'_j \psi_j(\mathbf{x}_i) + b \right) \geq 1 - \xi_i \\ & \xi \geq 0 ; \quad \beta \geq 0 ; \quad \|\beta\|_p \leq 1. \end{aligned} \quad (1)$$

The usage of kernels is permitted through its partially dualized form:

$$\begin{aligned} \min_{\beta} \max_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,l=1}^n \alpha_i \alpha_l y_i y_l \sum_{j=1}^m \beta_j k_j(\mathbf{x}_i, \mathbf{x}_l) \\ \text{s.t.} \quad & \forall_{i=1}^n : 0 \leq \alpha_i \leq C ; \quad \sum_{i=1}^n y_i \alpha_i = 0 ; \\ & \forall_{j=1}^m : \beta_j \geq 0 ; \quad \|\beta\|_p \leq 1. \end{aligned}$$

For details on the solution of this optimization problem and its kernelization we refer to the supplement and [21].

While prior work on MKL imposes a 1-norm constraint on the mixing coefficients to enforce sparse solutions lying on a standard simplex [14, 15, 32, 33], we employ a generalized  $\ell_p$ -norm constraint  $\|\beta\|_p \leq 1$  for  $p \geq 1$  as used in [20, 21]. The implications of this modification in the context of image concept classification will be discussed throughout this paper.

## 2.2 Kernel Target Alignment

The kernel alignment introduced by [34] measures the similarity of two matrices as a cosine angle of vectors un-

der the Frobenius product

$$\mathcal{A}(K_1, K_2) := \frac{\langle K_1, K_2 \rangle_F}{\|K_1\|_F \|K_2\|_F}, \quad (2)$$

It was argued in [35] that centering is required in order to correctly reflect the test errors from SVMs via kernel alignment. Centering in the corresponding feature spaces [36] can be achieved by taking the product  $HKH$ , with

$$H := I - \frac{1}{n} \mathbf{1} \mathbf{1}^\top, \quad (3)$$

$I$  is the identity matrix of size  $n$  and  $\mathbf{1}$  is the column vector with all ones. The centered kernel which achieves a perfect separation of two classes is proportional to  $\tilde{\mathbf{y}} \tilde{\mathbf{y}}^\top$ , where

$$\tilde{\mathbf{y}} = (\tilde{y}_i), \quad \tilde{y}_i := \begin{cases} \frac{1}{n_+} & y_i = +1 \\ -\frac{1}{n_-} & y_i = -1 \end{cases} \quad (4)$$

and  $n_+$  and  $n_-$  are the sizes of the positive and negative classes, respectively.

## 3 Empirical Evaluation

In this section, we evaluate  $\ell_p$ -norm MKL in real-world image categorization tasks, experimenting on the VOC2009 and ImageCLEF2010 data sets. We also provide insights on *when* and *why*  $\ell_p$ -norm MKL can help performance in image classification applications. The evaluation measure for both datasets is the average precision (AP) over all recall values based on the precision-recall (PR) curves.

### 3.1 Data Sets

We experiment on the following data sets:

**1. PASCAL2 VOC Challenge 2009** We use the official data set of the *PASCAL2 Visual Object Classes Challenge 2009* (VOC2009) [25], which consists of 13979 images. The use the official split into 3473 training, 3581 validation, and 6925 test examples as provided by the challenge organizers. The organizers also provided annotation of the 20 objects categories; note that an image can have multiple object annotations. The task is to solve 20 binary classification problems, i.e. predicting whether at

least one object from a class  $k$  is visible in the test image. Although the test labels are undisclosed, the more recent VOC datasets permit to evaluate AP scores on the test set via the challenge website (the number of allowed submissions per week being limited).

**2. ImageCLEF 2010 PhotoAnnotation** The ImageCLEF2010 PhotoAnnotation data set [26] consists of 8000 labeled training images taken from flickr and a test set with undisclosed labels. The images are annotated by 93 concept classes having highly variable concepts—they contain both well defined objects such as *lake, river, plants, trees, flowers*, as well as many rather ambiguously defined concepts such as *winter, boring, architecture, macro, artificial, motion blur*—however, those concepts might not always be connected to objects present in an image or captured by a bounding box. This makes it highly challenging for any recognition system. Unfortunately, there is currently no official way to obtain test set performance scores from the challenge organizers. Therefore, for this data set, we report on training set cross-validation performances only. As for VOC2009 we decompose the problem into 93 binary classification problems. Again, many concept classes are challenging to rank or classify by an object detection approach due to their inherent non-object nature. As for the previous dataset each image can be labeled with multiple concepts.

### 3.2 Image Features and Base Kernels

In all of our experiments we deploy 32 kernels capturing various aspects of the images. The kernels are inspired by the VOC 2007 winner [37] and our own experiences from our submissions to the VOC2009 and ImageCLEF2009 challenges. We can summarize the employed kernels by the following three types of basic features:

- Histogram over a bag of visual words over SIFT features (BoW-S), 15 kernels
- Histogram over a bag of visual words over color intensity histograms (BoW-C), 8 kernels
- Histogram of oriented gradients (HoG), 4 kernels
- Histogram of pixel color intensities (HoC), 5 kernels.

We used a higher fraction of bag-of-word-based features as we knew from our challenge submissions that

they have a better performance than global histogram features. The intention was, however, to use a variety of different feature types that have been proven to be effective on the above datasets in the past—but at the same time obeying memory limitations of maximally 25GB per job as required by computer facilities used in our experiments (we used a cluster of 23 nodes having in total 256 AMD64 CPUs and with memory limitations ranging in 32–96 GB RAM per node).

The above features are derived from histograms that contain *no* spatial information. We therefore enrich the respective representations by using spatial tilings  $1 \times 1$ ,  $3 \times 1$ ,  $2 \times 2$ ,  $4 \times 4$ ,  $8 \times 8$ , which correspond to single levels of the pyramidal approach [9] (this is for capturing the spatial context of an image). Furthermore, we apply a  $\chi^2$  kernel on top of the enriched histogram features, which is an established kernel for capturing histogram features [10]. The bandwidth of the  $\chi^2$  kernel is thereby heuristically chosen as the mean  $\chi^2$  distance over all pairs of training examples [38].

The BoW features were constructed in a standard way [39]: at first, the SIFT descriptors [40] were calculated on a regular grid with 6 pixel pitches for each image, learning a code book of size 4000 for the SIFT features and of size 900 for the color histograms by  $k$ -means clustering (with a random initialization). Finally, all SIFT descriptors were assigned to visual words (so-called *prototypes*) and then summarized into histograms within entire images or sub-regions. We computed the SIFT features over the following color combinations, which are inspired by the winners of the Pascal VOC 2008 challenge winners from the university of Amsterdam [41]: red-green-blue (RGB), normalized RGB, gray-opponentColor1-opponentColor2, and gray-normalized OpponentColor1-OpponentColor2; in addition, we also use a simple gray channel.

We computed the 15-dimensional local color histograms over the color combinations red-green-blue, gray-opponentColor1-opponentColor2, gray, and hue (the latter being weighted by the pixel value of the value component in the HSV color representation).

This means, for BoW-S, we considered five color channels with three spatial tilings each ( $1 \times 1$ ,  $3 \times 1$ , and  $2 \times 2$ ), resulting in 15 kernels; for BoW-C, we considered four color channels with two spatial tilings each ( $1 \times 1$  and  $3 \times 1$ ), resulting in 8 kernels.

The HoG features were computed by discretizing the orientation of the gradient vector at each pixel into 24 bins and then summarizing the discretized orientations into histograms within image regions [42]. Canny detectors [43] are used to discard contributions from pixels, around which the image is almost uniform. We computed them over the color combinations red-green-blue, gray-opponentColor1-opponentColor2, and gray, thereby using the two spatial tilings  $4 \times 4$  and  $8 \times 8$ . For the experiments we used four kernels: a product kernel created from the two kernels with the red-green-blue color combination but using different spatial tilings, another product kernel created in the same way but using the gray-opponentColor1-opponentColor2 color combination, and the two kernels using the gray channel alone (but differing in their spatial tiling).

The HoC features were constructed by discretizing pixel-wise color values and computing their 15 bin histograms within image regions. To this end, we used the color combinations red-green-blue, gray-opponentColor1-opponentColor2, and gray. For each color combination the spatial tilings  $2 \times 2$ ,  $3 \times 1$ , and  $4 \times 4$  were tried. In the experiments we deploy five kernels: a product kernel created from the three kernels with different spatial tilings with colors red-green-blue, a product kernel created from the three kernels with color combination gray-opponentColor1-opponentColor2, and the three kernels using the gray channel alone (differing in their spatial tiling).

Note that building a product kernel out of  $\chi^2$  kernels boils down to concatenating feature blocks (but using a separate kernel width for each feature block). The intention here was to use single kernels at separate spatial tilings for the weaker features (for problems depending on a certain tiling resolution) and combined kernels with all spatial tilings merged into one kernel to keep the memory requirements low and let the algorithms select the best choice.

In practice, the normalization of kernels is as important for MKL as the normalization of features is for training regularized linear or single-kernel models. This is owed to the bias introduced by the regularization: optimal feature / kernel weights are requested to be small, implying a bias to towards excessively up-scaled kernels. In general, there are several ways of normalizing kernel functions. We apply the following normalization method, pro-

posed in [44, 45] and entitled *multiplicative normalization* in [21]; on the feature-space level this normalization corresponds to rescaling training examples to unit variance,

$$K \leftarrow \frac{1}{n} \text{tr}(K) - \frac{1}{n^2} \mathbf{1}^\top K \mathbf{1}. \quad (5)$$

### 3.3 Experimental Setup

We treat the multi-label data set as binary classification problems, that is, for each object category we trained a one-vs.-rest classifier. Multiple labels per image render multi-class methods inapplicable as these require mutually exclusive labels for the images. The respective SVMs are trained using the Shogun toolbox [46]. In order to shed light on the nature of the presented techniques from a statistical viewpoint, we first pooled all labeled data and then created 20 random cross-validation splits for VOC2009 and 12 splits for the larger dataset ImageCLEF2010.

For each of the 12 or 20 splits, the training images were used for learning the classifiers, while the SVM/MKL regularization parameter  $C$  and the norm parameter  $p$  were chosen based on the maximal AP score on the validation images. Thereby, the regularization constant  $C$  is optimized by class-wise grid search over  $C \in \{10^i \mid i = -1, -0.5, 0, 0.5, 1\}$ . Preliminary runs indicated that this way the optimal solutions are attained inside the grid. Note that for  $p = \infty$  the  $\ell_p$ -norm MKL boils down to a simple SVM using a uniform kernel combination (subsequently called sum-kernel SVM). In our experiments, we used the average kernel SVM instead of the sum-kernel one. This is no limitation in this as both lead to identical result for an appropriate choice of the SVM regularization parameter.

For a rigorous evaluation, we would have to construct a separate codebook for each cross validation split. However, creating codebooks and assigning descriptors to visual words is a time-consuming process. Therefore, in our experiments we resort to the common practice of using a single codebook created from all training images contained in the official split. Although this could result in a slight overestimation of the AP scores, this affects all methods equally and does not favor any classification method more than another—our focus lies on a *relative* comparison of the different classification methods; there-

fore there is no loss in exploiting this computational short-cut.

### 3.4 Results

In this section we report on the empirical results achieved by  $\ell_p$ -norm MKL in our visual object recognition experiments.

**VOC 2009** Table 2 shows the AP scores attained on the official test split of the VOC2009 data set (scores obtained by evaluation via the challenge website). The class-wise optimal regularization constant has been selected by cross-validation-based model selection on the training data set. We can observe that non-sparse MKL outperforms the baselines  $\ell_1$ -MKL and the sum-kernel SVM in this sound evaluation setup. We also report on the cross-validation performance achieved on the training data set (Table 1). Comparing the two results, one can observe a small overestimation for the cross-validation approach (for the reasons argued in Section 3.3)—however, the amount by which this happens is equal for all methods; in particular, the ranking of the compared methods (SVM versus  $\ell_p$ -norm MKL for various values of  $p$ ) is preserved for the average over all classes and most of the classes (exceptions are the bottle and bird class); this shows the reliability of the cross-validation-based evaluation method in practice. Note that the observed variance in the AP measure across concepts can be explained in part by the variations in the label distributions across concepts and cross-validation splits. Unlike for the AUC measure, the average score of the AP measure under randomly ranked images depends on the ratio of positive and negative labeled samples.

A reason why the bottle class shows such a strong deviation towards sparse methods could be the varying but often small fraction of image area covered by bottles leading to overfitting when using spatial tilings.

We can also remark that  $\ell_{1.333}$ -norm achieves the best result of all compared methods on the VOC dataset, slightly followed by  $\ell_{1.125}$ -norm MKL. To evaluate the statistical significance of our findings, we perform a Wilcoxon signed-rank test for the cross-validation-based results (see Table 1; significant results are marked in bold-face). We find that in 15 out of the 20 classes the optimal result is achieved by truly non-sparse  $\ell_p$ -norm MKL

(which means  $p \in ]1, \infty[$ ), thus outperforming the baseline significantly.

**ImageCLEF** Table 4 shows the AP scores averaged over all classes achieved on the ImageCLEF2010 data set. We observe that the best result is achieved by the non-sparse  $\ell_p$ -norm MKL algorithms with norm parameters  $p = 1.125$  and  $p = 1.333$ . The detailed results for all 93 classes are shown in the supplemental material (see B.1 and B.2). We can see from the detailed results that in 37 out of the 93 classes the optimal result attained by non-sparse  $\ell_p$ -norm MKL was significantly better than the sum kernel according to a Wilcoxon signed-rank test.

We also show the results for optimizing the norm parameter  $p$  class-wise (see Table 5). We can see from the table that optimizing the  $\ell_p$ -norm class-wise is beneficial: selecting the best  $p \in ]1, \infty[$  class-wise, the result is increased to an AP of 39.70. Also including  $\ell_1$ -norm MKL in the candidate set, the performance can even be leveraged to 39.82—this is 0.7 AP better than the result for the vanilla sum-kernel SVM. Also including the latter to the set of model, the AP score only merely increases by 0.03 AP points. We conclude that optimizing the norm parameter  $p$  class-wise can improve performance; however, one can rely on  $\ell_p$ -norm MKL alone without the need to additionally include the sum-kernel-SVM to the set of models. Tables 1 and 2 show that the gain in performance for MKL varies considerably on the actual concept class. Notice that these observations are confirmed by the results presented in Tables B.1 and B.2, see supplemental material for details.

### 3.5 Analysis and Interpretation

We now analyze the kernel set in an explorative manner; to this end, our methodological tools are the following

1. Pairwise kernel alignment scores (KA)
2. Centered kernel-target alignment scores (KTA).

#### 3.5.1 Analysis of the Chosen Kernel Set

To start with, we computed the pairwise kernel alignment scores of the 32 base kernels: they are shown in Fig. 1. We recall that the kernels can be classified into the following groups: Kernels 1–15 and 16–23 employ BoW-S and

Table 1: Average AP scores on the VOC2009 data set with AP scores computed by cross-validation on the training set. Bold faces show the best method and all other ones that are not statistical-significantly worse.

Norm	Average	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus
$\ell_1$	54.94 $\pm$ 12.3	<b>84.84</b> $\pm$ 5.86	55.35 $\pm$ 10.5	59.38 $\pm$ 10.1	66.83 $\pm$ 12.4	25.91 $\pm$ 10.2	<b>71.15</b> $\pm$ 23.2
$\ell_{1.125}$	57.07 $\pm$ 12.7	<b>84.82</b> $\pm$ 5.91	<b>57.25</b> $\pm$ 10.6	62.4 $\pm$ 9.13	<b>67.89</b> $\pm$ 12.8	<b>27.88</b> $\pm$ 9.91	<b>71.7</b> $\pm$ 22.8
$\ell_{1.333}$	57.2 $\pm$ 12.8	<b>84.51</b> $\pm$ 6.27	<b>57.41</b> $\pm$ 10.8	<b>62.75</b> $\pm$ 9.07	<b>67.99</b> $\pm$ 13	<b>27.44</b> $\pm$ 9.77	<b>71.33</b> $\pm$ 23.1
$\ell_2$	56.53 $\pm$ 12.8	84.12 $\pm$ 5.92	56.89 $\pm$ 10.9	<b>62.53</b> $\pm$ 8.9	67.69 $\pm$ 13	26.68 $\pm$ 9.94	70.33 $\pm$ 22.3
$\ell_\infty$	56.08 $\pm$ 12.7	83.67 $\pm$ 5.99	56.09 $\pm$ 10.9	61.91 $\pm$ 8.81	67.52 $\pm$ 12.9	26.5 $\pm$ 9.5	70.13 $\pm$ 22.2
Norm	Car	Cat	Chair	Cow	Diningtable	Dog	Horse
$\ell_1$	54.54 $\pm$ 7.33	59.5 $\pm$ 8.22	53.3 $\pm$ 11.7	23.13 $\pm$ 13.2	<b>48.51</b> $\pm$ 19.9	41.72 $\pm$ 9.44	57.67 $\pm$ 12.2
$\ell_{1.125}$	<b>56.59</b> $\pm$ 8.93	<b>61.59</b> $\pm$ 8.26	<b>54.3</b> $\pm$ 12.1	<b>29.59</b> $\pm$ 16.2	<b>49.32</b> $\pm$ 19.5	<b>45.57</b> $\pm$ 10.6	<b>59.4</b> $\pm$ 12.2
$\ell_{1.333}$	<b>56.75</b> $\pm$ 9.28	<b>61.74</b> $\pm$ 8.41	<b>54.25</b> $\pm$ 12.3	<b>29.89</b> $\pm$ 15.8	48.4 $\pm$ 19.3	<b>45.85</b> $\pm$ 10.9	<b>59.4</b> $\pm$ 11.9
$\ell_2$	55.92 $\pm$ 9.49	61.39 $\pm$ 8.37	<b>53.85</b> $\pm$ 12.4	28.39 $\pm$ 16.2	47 $\pm$ 18.7	45.14 $\pm$ 10.8	58.61 $\pm$ 11.9
$\ell_\infty$	55.58 $\pm$ 9.47	<b>61.25</b> $\pm$ 8.28	53.13 $\pm$ 12.4	27.56 $\pm$ 16.2	46.29 $\pm$ 18.8	44.63 $\pm$ 10.6	58.32 $\pm$ 11.7
Norm	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	Tvmonitor
$\ell_1$	55 $\pm$ 13.2	81.32 $\pm$ 9.49	35.14 $\pm$ 13.4	38.13 $\pm$ 19.2	<b>48.15</b> $\pm$ 11.8	<b>75.33</b> $\pm$ 14.1	63.97 $\pm$ 10.2
$\ell_{1.125}$	<b>57.66</b> $\pm$ 13.1	<b>82.18</b> $\pm$ 9.3	39.05 $\pm$ 14.9	<b>43.65</b> $\pm$ 20.5	<b>48.72</b> $\pm$ 13	<b>75.79</b> $\pm$ 14.4	<b>65.99</b> $\pm$ 9.83
$\ell_{1.333}$	<b>57.57</b> $\pm$ 13	<b>82.27</b> $\pm$ 9.29	<b>39.7</b> $\pm$ 14.6	<b>46.28</b> $\pm$ 23.9	<b>48.76</b> $\pm$ 11.9	<b>75.75</b> $\pm$ 14.3	<b>66.07</b> $\pm$ 9.59
$\ell_2$	<b>56.9</b> $\pm$ 13.2	82.19 $\pm$ 9.3	38.97 $\pm$ 14.8	<b>45.88</b> $\pm$ 24	47.29 $\pm$ 11.7	<b>75.29</b> $\pm$ 14.5	<b>65.55</b> $\pm$ 10.1
$\ell_\infty$	<b>56.45</b> $\pm$ 13.1	82 $\pm$ 9.37	38.46 $\pm$ 14.1	<b>45.93</b> $\pm$ 24	46.08 $\pm$ 11.8	74.89 $\pm$ 14.5	<b>65.19</b> $\pm$ 10.2

Table 2: AP scores attained on the VOC2009 test data, obtained on request from the challenge organizers. Best methods are marked boldface.

	average	aeroplane	bicycle	bird	boat	bottle	bus	car
$\ell_1$	54.58	<b>81.13</b>	54.52	56.14	62.44	<b>28.10</b>	<b>68.92</b>	52.33
$\ell_{1.125}$	56.43	81.01	56.36	58.49	62.84	25.75	68.22	55.71
$\ell_{1.333}$	<b>56.70</b>	80.77	<b>56.79</b>	<b>58.88</b>	63.11	25.26	67.80	<b>55.98</b>
$\ell_2$	56.34	80.41	56.34	58.72	<b>63.13</b>	24.55	67.70	55.54
$\ell_\infty$	55.85	79.80	55.68	58.32	62.76	24.23	67.79	55.38
	cat	chair	cow	diningtable	dog	horse	motorbike	
$\ell_1$	55.50	52.22	36.17	45.84	41.90	61.90	57.58	
$\ell_{1.125}$	57.79	53.66	40.77	<b>48.40</b>	46.36	<b>63.10</b>	60.89	
$\ell_{1.333}$	<b>58.00</b>	<b>53.87</b>	<b>43.14</b>	48.17	46.54	63.08	<b>61.28</b>	
$\ell_2$	57.98	53.47	40.95	48.07	<b>46.59</b>	63.02	60.91	
$\ell_\infty$	57.30	53.07	39.74	47.27	45.87	62.49	60.55	
	person	pottedplant	sheep	sofa	train	tvmonitor		
$\ell_1$	81.73	31.57	36.68	45.72	<b>80.52</b>	61.41		
$\ell_{1.125}$	82.65	<b>34.61</b>	41.91	<b>46.59</b>	80.13	63.51		
$\ell_{1.333}$	<b>82.72</b>	34.60	44.14	46.42	79.93	<b>63.60</b>		
$\ell_2$	82.52	33.40	<b>44.81</b>	45.98	79.53	63.26		
$\ell_\infty$	82.20	32.76	44.15	45.69	79.03	63.00		

Table 3: Average AP scores on the VOC2009 data set with norm parameter  $p$  class-wise optimized over AP scores on the training set. We report on test set scores obtained on request from the challenge organizers.

$\infty$	$\{1, \infty\}$	$\{1.125, 1.333, 2\}$	$\{1.125, 1.333, 2, \infty\}$	$\{1, 1.125, 1.333, 2\}$	all norms from the left
55.85	55.94	56.75	56.76	56.75	56.76

Table 4: Average AP scores obtained on the ImageCLEF2010 data set with  $p$  fixed over the classes and AP scores computed by cross-validation on the training set.

$\ell_p$ -Norm	1	1.125	1.333	2	$\infty$
	$37.32 \pm 5.87$	$39.51 \pm 6.67$	$39.48 \pm 6.66$	$39.13 \pm 6.62$	$39.11 \pm 6.68$

BoW-C features, respectively; Kernels 24 to 27 are product kernels associated with the HoG and HoC features; Kernels 28–30 deploy HoC, and, finally, Kernels 31–32 are based on HoG features over the gray channel. We see from the block-diagonal structure that features that are of the same type (but are generated for different parameter values, color channels, or spatial tilings) are strongly correlated. Furthermore the BoW-S kernels (Kernels 1–15) are weakly correlated with the BoW-C kernels (Kernels 16–23). Both, the BoW-S and HoG kernels (Kernels 24–25, 31–32) use gradients and therefore are moderately correlated; the same holds for the BoW-C and HoC kernel groups (Kernels 26–30). This corresponds to our original intention to have a broad range of feature types which are, however, useful for the task at hand. The principle usefulness of our feature set can be seen a posteriori from the fact that  $\ell_1$ -MKL achieves the worst performance of all methods included in the comparison while the sum-kernel SVM performs moderately well. Clearly, a higher fraction of noise kernels would further harm the sum-kernel SVM and favor the sparse MKL instead (we investigate the impact of noise kernels on the performance of  $\ell_p$ -norm MKL in an experiment on controlled, artificial data; this is presented in the supplemental material).

Based on the observation that the BoW-S kernel subset shows high KTA scores, we also evaluated the performance restricted to the 15 BoW-S kernels only. Unsurprisingly, this setup favors the sum-kernel SVM, which achieves higher results on VOC2009 for most classes; compared to  $\ell_p$ -norm MKL using all 32 classes, the sum-kernel SVM restricted to 15 classes achieves slightly better AP scores for 11 classes, but also slightly worse for 9 classes. Furthermore, the sum kernel SVM,  $\ell_2$ -MKL,

and  $\ell_{1.333}$ -MKL were on par with differences fairly below 0.01 AP. This is again not surprising as the kernels from the BoW-S kernel set are strongly correlated with each other for the VOC data which can be seen in the top left image in Fig. 1. For the ImageCLEF data we observed a quite different picture: the sum-kernel SVM restricted to the 15 BoW-S kernels performed significantly worse, when, again, being compared to non-sparse  $\ell_p$ -norm MKL using all 32 kernels. To achieve top state-of-the-art performance, one could optimize the scores for both datasets by considering the class-wise maxima over learning methods *and* kernel sets. However, since the intention here is not to win a challenge but a relative comparison of models, giving insights in the nature of the methods—we therefore discard the time-consuming optimization over the kernel subsets.

From the above analysis, the question arises why restricting the kernel set to the 15 BoW-S kernels affects the performance of the compared methods differently, for the VOC2009 and ImageCLEF2010 data sets. This can be explained by comparing the KA/KTA scores of the kernels attained on VOC and on ImageCLEF (see Fig. 1 (RIGHT)): for the ImageCLEF data set the KTA scores are substantially more spread along all kernels; there is neither a dominance of the BoW-S subset in the KTA scores nor a particularly strong correlation within the BoW-S subset in the KA scores. We attribute this to the less object-based and more ambiguous nature of many of the concepts contained in the ImageCLEF data set. Furthermore, the KA scores for the ImageCLEF data (see Fig. 1 (LEFT)) show that this dataset exhibits a higher variance among kernels—this is because the correlations between all kinds of kernels are weaker for the Image-



Table 5: Average AP scores obtained on the ImageCLEF2010 data set with norm parameter  $p$  class-wise optimized and AP scores computed by cross-validation on the training set.

$\infty$	$\{1, \infty\}$	$\{1.125, 1.333, 2\}$	$\{1.125, 1.333, 2, \infty\}$	$\{1, 1.125, 1.333, 2\}$	all norms from the left
$39.11 \pm 6.68$	$39.33 \pm 6.71$	$39.70 \pm 6.80$	$39.74 \pm 6.85$	$39.82 \pm 6.82$	$39.85 \pm 6.88$

CLEF data.

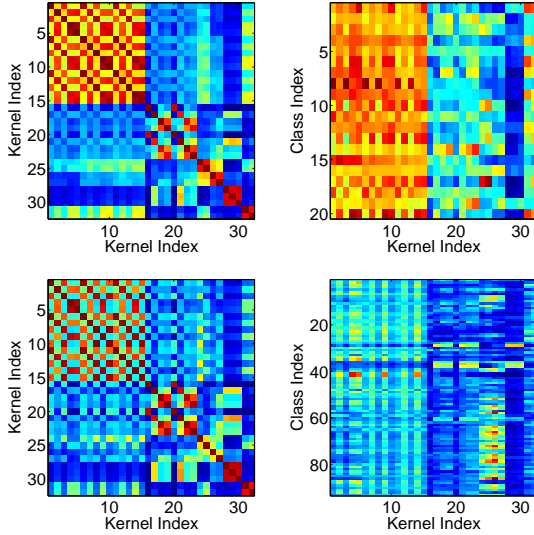


Figure 1: Similarity of the kernels for the VOC2009 (TOP) and ImageCLEF2010 (BOTTOM) data sets in terms of pairwise kernel alignments (LEFT) and kernel target alignments (RIGHT), respectively. In both data sets, five groups can be identified: 'BoW-S' (Kernels 1–15), 'BoW-C' (Kernels 16–23), 'products of HoG and HoC kernels' (Kernels 24–27, 'HoC single' (Kernels 28–30), and 'HoG single' (Kernels 31–32).

Therefore, because of this non-uniformity in the spread of the information content among the kernels, we can conclude that indeed our experimental setting falls into the situation where non-sparse MKL can outperform the baseline procedures (again, see supplemental material). For example, the BoW features are more informative than HoG and HoC, and thus the uniform-sum-kernel-SVM is suboptimal. On the other hand, because of the fact that typical image features are only moderately informative, HoG and HoC still convey a certain amount of complementary information—this is what allows the perfor-

mance gains reported in Tables 1 and 4.

Note that we class-wise normalized the KTA scores to sum to one. This is because we are rather interested in a comparison of the relative contributions of the particular kernels than in their absolute information content, which anyway can be more precisely derived from the AP scores already reported in Tables 1 and 4. Furthermore, note that we consider *centered* KA and KTA scores, since it was argued in [35] that only those correctly reflect the test errors attained by established learners such as SVMs.

### 3.5.2 The Role of the Choice of $\ell_p$ -norm

Next, we turn to the interpretation of the norm parameter  $p$  in our algorithm. We observe a big gap in performance between  $\ell_{1.125}$ -norm MKL and the sparse  $\ell_1$ -norm MKL. The reason is that for  $p > 1$  MKL is reluctant to set kernel weights to zero, as can be seen from Figure 2. In contrast,  $\ell_1$ -norm MKL eliminates 62.5% of the kernels from the working set. The difference between the  $\ell_p$ -norms for  $p > 1$  lies solely in the ratio by which the less informative kernels are down-weighted—they are never assigned with true zeros.

However, as proved in [21], in the computational optimum, the kernel weights are accessed by the MKL algorithm via the information content of the particular kernels given by a  $\ell_p$ -norm-dependent formula (see Eq. (8); this will be discussed in detail in Section 4.1). We mention at this point that the kernel weights all converge to the same, uniform value for  $p \rightarrow \infty$ . We can confirm these theoretical findings empirically: the histograms of the kernel weights shown in Fig. 2 clearly indicate an increasing uniformity in the distribution of kernel weights when letting  $p \rightarrow \infty$ . Higher values of  $p$  thus cause the weight distribution to shift away from zero and become slanted to the right while smaller ones tend to increase its skewness to the left.

Selection of the  $\ell_p$ -norm permits to tune the strength of the regularization of the learning of kernel weights. In this

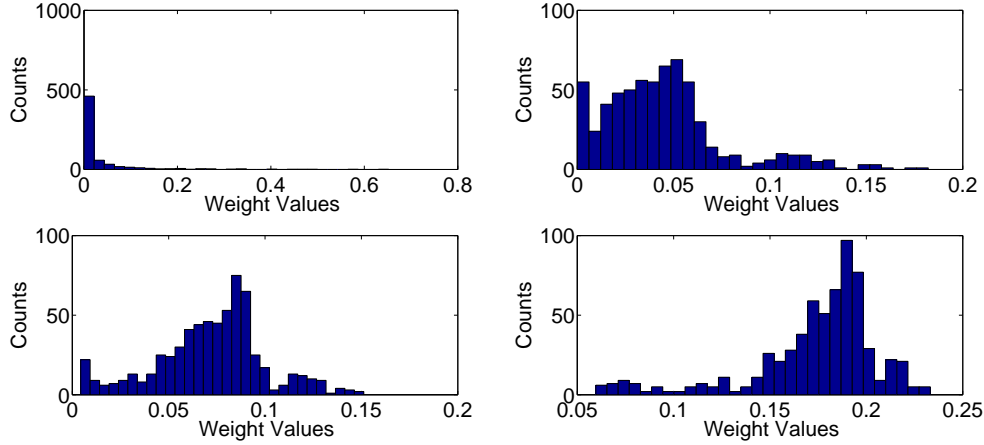


Figure 2: Histogram of kernel weights as output by  $\ell_p$ -norm MKL for the various classes on the VOC2009 data set (32 kernels  $\times$  20 classes, resulting in 640 values):  $\ell_1$ -norm (TOP LEFT),  $\ell_{1.125}$ -norm (TOP RIGHT),  $\ell_{1.333}$ -norm (BOTTOM LEFT), and  $\ell_2$ -norm (BOTTOM RIGHT).

sense the sum-kernel SVM clearly is an extreme, namely fixing the kernel weights, obtained when letting  $p \rightarrow \infty$ . The sparse MKL marks another extreme case:  $\ell_p$ -norms with  $p$  below 1 loose the convexity property so that  $p = 1$  is the maximally sparse choice preserving convexity at the same time. Sparsity can be interpreted here that only a few kernels are selected which are considered most informative according to the optimization objective. Thus, the  $\ell_p$ -norm acts as a prior parameter for how much we trust in the informativeness of a kernel. In conclusion, this interpretation justifies the usage of  $\ell_p$ -norm outside the existing choices  $\ell_1$  and  $\ell_2$ . The fact that the sum-kernel SVM is a reasonable choice in the context of image annotation will be discussed further in Section 4.1.

Our empirical findings on ImageCLEF and VOC seem to contradict previous ones about the usefulness of MKL reported in the literature, where  $\ell_1$  is frequently to be outperformed by a simple sum-kernel SVM (for example, see [30])—however, in these studies the sum-kernel SVM is compared to  $\ell_1$ -norm or  $\ell_2$ -norm MKL only. In fact, our results *confirm* these findings:  $\ell_1$ -norm MKL is outperformed by the sum-kernel SVM in all of our experiments. Nevertheless, in this paper, we show that by using the more general  $\ell_p$ -norm regularization, the prediction accuracy of MKL can be considerably leveraged, even clearly outperforming the sum-kernel SVM, which

has been shown to be a tough competitor in the past [12]. But of course also the simpler sum-kernel SVM also has its advantage, although on the computational side only: in our experiments it was about a factor of ten faster than its MKL competitors. Further information about runtimes of MKL algorithms compared to sum kernel SVMs can be taken from [47].

### 3.5.3 Remarks for Particular Concepts

Finally, we show images from classes where MKL helps performance and discuss relationships to kernel weights. We have seen above that the sparsity-inducing  $\ell_1$ -norm MKL clearly outperforms all other methods on the *bottle* class (see Table 2). Fig. 3 shows two typical highly ranked images and the corresponding kernel weights as output by  $\ell_1$ -norm (LEFT) and  $\ell_{1.333}$ -norm MKL (RIGHT), respectively, on the bottle class. We observe that  $\ell_1$ -norm MKL tends to rank highly party and people group scenes. We conjecture that this has two reasons: first, many people group and party scenes come along with co-occurring bottles. Second, people group scenes have similar gradient distributions to images of large upright standing bottles sharing many dominant vertical lines and a thinner head section—see the left- and right-hand images in Fig. 3. Sparse  $\ell_1$ -norm MKL strongly focuses on the

dominant HoG product kernel, which is able to capture the aforementioned special gradient distributions, giving small weights to two HoC product kernels and almost completely discarding all other kernels.

Next, we turn to the *cow* class, for which we have seen above that  $\ell_{1.333}$ -norm MKL outperforms all other methods clearly. Fig. 4 shows a typical high-ranked image of that class and also the corresponding kernel weights as output by  $\ell_1$ -norm (LEFT) and  $\ell_{1.333}$ -norm (RIGHT) MKL, respectively. We observe that  $\ell_1$ -MKL focuses on the two HoC product kernels; this is justified by typical cow images having green grass in the background. This allows the HoC kernels to easily distinguish the cow images from the indoor and vehicle classes such as *car* or *sofa*. However, horse and sheep images have such a green background, too. They differ in sheep usually being black-white, and horses having a brown-black color bias (in VOC data); cows have rather variable colors. Here, we observe that the rather complex yet somewhat color-based BoW-C and BoW-S features help performance—it is also those kernels that are selected by the non-sparse  $\ell_{1.333}$ -MKL, which is the best performing model on those classes. In contrast, the sum-kernel SVM suffers from including the five gray-channel-based features, which are hardly useful for the horse and sheep classes and mostly introduce additional noise. MKL (all variants) succeed in identifying those kernels and assign those kernels with low weights.

## 4 Discussion

In the previous section we presented empirical evidence that  $\ell_p$ -norm MKL considerably can help performance in visual image categorization tasks. We also observed that the gain is class-specific and limited for some classes when compared to the sum-kernel SVM, see again Tables 1 and 2 as well as Tables B.1, B.2 in the supplemental material. In this section, we aim to shed light on the reasons of this behavior, in particular discussing strengths of the average kernel in Section 4.1, trade-off effects in Section 4.2 and strengths of MKL in Section 4.3. Since these scenarios are based on statistical properties of kernels which can be observed in concept recognition tasks within computer vision we expect the results to be transferable to other algorithms which learn linear models over

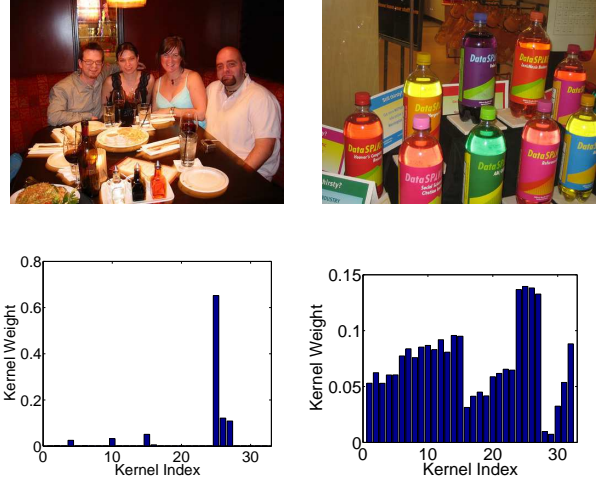


Figure 3: Images of typical highly ranked bottle images and kernel weights from  $\ell_1$ -MKL (left) and  $\ell_{1.333}$ -MKL (right).

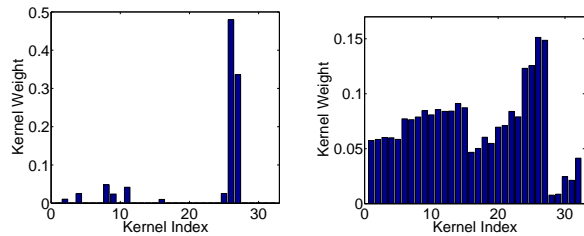


Figure 4: Images of a typical highly ranked cow image and kernel weights from  $\ell_1$ -MKL (left) and  $\ell_{1.333}$ -MKL (right).

kernels such as [28, 29].

#### 4.1 One Argument For the Sum Kernel: Randomness in Feature Extraction

We would like to draw attention to one aspect present in BoW features, namely the amount of randomness induced by the visual word generation stage acting as noise with respect to kernel selection procedures.

**Experimental setup** We consider the following experiment, similar to the one undertaken in [30]: we compute a BoW kernel ten times each time using the same local features, identical spatial pyramid tilings, and identical kernel functions; the only difference between subsequent repetitions of the experiment lies in the randomness involved in the generation of the codebook of visual words. Note that we use SIFT features over the gray channel that are densely sampled over a grid of step size six, 512 visual words (for computational feasibility of the clustering), and a  $\chi^2$  kernel. This procedure results in ten kernels that only differ in the randomness stemming from the codebook generation. We then compare the performance of the sum-kernel SVM built from the ten kernels to the one of the best single-kernel SVM determined by cross-validation-based model selection.

In contrast to [30] we try *two* codebook generation procedures, which differ by their intrinsic amount of randomness: first, we deploy *k*-means clustering, with random initialization of the centers and a bootstrap-like selection of the best initialization (similar to the option 'cluster' in MATLAB's *k*-means routine). Second, we deploy *extremely randomized clustering forests* (ERCF) [48, 49], that are, ensembles of randomized trees—the latter procedure involves a considerably higher amount of randomization compared to *k*-means.

**Results** The results are shown in Table 6. For both clustering procedures, we observe that the sum-kernel SVM outperforms the best single-kernel SVM. In particular, this confirms earlier findings of [30] carried out for *k*-means-based clustering. We also observe that the difference between the sum-kernel SVM and the best single-kernel SVM is much more pronounced for ERCF-based kernels—we conclude that this stems from a higher amount of randomness is involved in the ERCF clustering method when compared to conventional *k*-means. The

standard deviations of the kernels in Table 6 confirm this conclusion. For each class we computed the conditional standard deviation

$$\text{std}(K \mid y_i = y_j) + \text{std}(K \mid y_i \neq y_j) \quad (6)$$

averaged over all classes. The usage of a conditional variance estimator is justified because the ideal similarity in kernel target alignment (cf. equation (4)) does have a variance over the kernel as a whole however the conditional deviations in equation (6) would be zero for the ideal kernel. Similarly, the fundamental MKL optimization formula (8) relies on a statistic based on the two conditional kernels used in formula (6). Finally, ERCF clustering uses label information. Therefore averaging the class-wise conditional standard deviations over all classes is not expected to be identical to the standard deviation of the whole kernel.

We observe in Table 6 that the standard deviations are lower for the sum kernels. Comparing ERCF and *k*-means shows that the former not only exhibits larger absolute standard deviations but also greater differences between single-best and sum-kernel as well as larger differences in AP scores.

We can thus postulate that the reason for the superior performance of the sum-kernel SVM stems from averaging out the randomness contained in the BoW kernels (stemming from the visual-word generation). This can be explained by the fact that averaging is a way of reducing the variance in the predictors/models [50]. We can also remark that such variance reduction effects can also be observed when averaging BoW kernels with varying color combinations or other parameters; this stems from the randomness induced by the visual word generation.

Note that in the above experimental setup each kernel uses the *same* information provided via the local features. Consequently, the best we can do is *averaging*—learning kernel weights in such a scenario is likely to suffer from overfitting to the noise contained in the kernels and can only decrease performance.

To further analyze this, we recall that, in the computational optimum, the information content of a kernel is measured by  $\ell_p$ -norm MKL via the following quantity, as

Method	Best Single Kernel	Sum Kernel
VOC09, k-Means	AP: 44.42 ± 12.82	<b>45.84</b> ± 12.94
VOC09, k-Means	Std: <b>30.81</b>	30.74
VOC09, ERCF	AP: 42.60 ± 12.50	<b>47.49</b> ± 12.89
VOC09, ERCF	Std: <b>38.12</b>	37.89
ImageCLEF, k-Means	AP: 31.09 ± 5.56	<b>31.73</b> ± 5.57
ImageCLEF, k-Means	Std: <b>30.51</b>	30.50
ImageCLEF, ERCF	AP: 29.91 ± 5.39	<b>32.77</b> ± 5.93
ImageCLEF, ERCF	Std: <b>38.58</b>	38.10

Table 6: AP Scores and standard deviations showing amount of randomness in feature extraction: results from repeated computations of BoW Kernels with randomly initialized codebooks

proved in [21]:

$$\beta \propto \|w\|_2^{\frac{2}{p+1}} = \left( \sum_{i,j} \alpha_i y_i K_{ij} \alpha_j y_j \right)^{\frac{2}{p+1}}. \quad (7)$$

In this paper we deliver a novel interpretation of the above quantity; to this end, we decompose the right-hand term into two terms as follows:

$$\sum_{i,j} \alpha_i y_i K_{ij} \alpha_j y_j = \sum_{i,j|y_i=y_j} \alpha_i K_{ij} \alpha_j - \sum_{i,j|y_i \neq y_j} \alpha_i K_{ij} \alpha_j.$$

The above term can be interpreted as a difference of the support-vector-weighted sub-kernel restricted to consistent labels *and* the support-vector-weighted sub-kernel over the opposing labels. Equation 7 thus can be rewritten as

$$\beta \propto \left( \sum_{i,j|y_i=y_j} \alpha_i K_{ij} \alpha_j - \sum_{i,j|y_i \neq y_j} \alpha_i K_{ij} \alpha_j \right)^{\frac{2}{p+1}}. \quad (8)$$

Thus, we observe that random influences in the features combined with overfitting support vectors can suggest a falsely high information content in this measure for *some* kernels. SVMs do overfit on BoW features. Using the scores attained on the training data subset we can observe that many classes are deceptive-perfectly predicted with AP scores fairly above 0.9. At this point, non-sparse  $\ell_{p>1}$ -norm MKL offers a parameter  $p$  for regularizing the kernel weights—thus hardening the algorithm to become robust against random noise, yet permitting to use some degree of information given by Equation (8).

[30] reported in accordance to our idea about overfitting of SVMs that  $\ell_2$ -MKL and  $\ell_1$ -MKL show no gain in such a scenario while  $\ell_1$ -MKL even reduces performance for some datasets. This result is not surprising as the overly sparse  $\ell_1$ -MKL has a stronger tendency to overfit to the randomness contained in the kernels / feature generation. The observed amount of randomness in the state-of-the-art BoW features could be an explanation why the sum-kernel SVM has shown to be a quite hard-to-beat competitor for semantic concept classification and ranking problems.

## 4.2 MKL and Prior Knowledge

For solving a learning problem, there is nothing more valuable than *prior knowledge*. Our empirical findings on the VOC2009 and ImageCLEF09 data sets suggested that our experimental setup was actually biased towards the sum-kernel SVM via usage of prior knowledge when choosing the set of kernels / image features. We deployed kernels based on four features types: BoW-S, BoW-C, HoC and HoG. However, the *number* of kernels taken from each feature type is not equal. Based on our experience with the VOC and ImageCLEF challenges we used a higher fraction of BoW kernels and less kernels of other types such as histograms of colors or gradients because we already knew that BoW kernels have superior performance.

To investigate to what extend our choice of kernels introduces a bias towards the sum-kernel SVM, we also performed another experiment, where we deployed a higher fraction of weaker kernels for VOC2009. The difference to our previous experiments lies in that we summarized the 15 BOW-S kernels in 5 product kernels reducing the number of kernels from 32 to 22. The results are given in Table 7; when compared to the results of the original 32-kernel experiment (shown in Table 1), we observe that the AP scores are in average about 4 points smaller. This can be attributed to the fraction of weak kernels being higher as in the original experiment; consequently, the gain from using ( $\ell_{1.333}$ -norm) MKL compared to the sum-kernel SVM is now more pronounced: over 2 AP points—again, this can be explained by the higher fraction of weak (i.e., noisy) kernels in the working set (this effect is also confirmed in the toy experiment carried out in supplemental material: there, we see that MKL becomes more bene-

Class / $\ell_p$ -norm	1.333	$\infty$
Aeroplane	<b>77.82</b> $\pm$ 7.701	76.28 $\pm$ 8.168
Bicycle	<b>50.75</b> $\pm$ 11.06	46.39 $\pm$ 12.37
Bird	<b>57.7</b> $\pm$ 8.451	55.09 $\pm$ 8.224
Boat	<b>62.8</b> $\pm$ 13.29	60.9 $\pm$ 14.01
Bottle	<b>26.14</b> $\pm$ 9.274	25.05 $\pm$ 9.213
Bus	<b>68.15</b> $\pm$ 22.55	67.24 $\pm$ 22.8
Car	<b>51.72</b> $\pm$ 8.822	49.51 $\pm$ 9.447
Cat	<b>56.69</b> $\pm$ 9.103	55.55 $\pm$ 9.317
Chair	<b>51.67</b> $\pm$ 12.24	49.85 $\pm$ 12
Cow	<b>25.33</b> $\pm$ 13.8	22.22 $\pm$ 12.41
Diningtable	<b>45.91</b> $\pm$ 19.63	42.96 $\pm$ 20.17
Dog	<b>41.22</b> $\pm$ 10.14	39.04 $\pm$ 9.565
Horse	<b>52.45</b> $\pm$ 13.41	50.01 $\pm$ 13.88
Motorbike	<b>54.37</b> $\pm$ 12.91	52.63 $\pm$ 12.66
Person	<b>80.12</b> $\pm$ 10.13	79.17 $\pm$ 10.51
Pottedplant	<b>35.69</b> $\pm$ 13.37	34.6 $\pm$ 14.09
Sheep	<b>37.05</b> $\pm$ 18.04	34.65 $\pm$ 18.68
Sofa	<b>41.15</b> $\pm$ 11.21	37.88 $\pm$ 11.11
Train	<b>70.03</b> $\pm$ 15.67	67.87 $\pm$ 16.37
Tvmonitor	<b>59.88</b> $\pm$ 10.66	57.77 $\pm$ 10.91
Average	<b>52.33</b> $\pm$ 12.57	50.23 $\pm$ 12.79

Table 7: MKL versus Prior Knowledge: AP Scores with a smaller fraction of well scoring kernels

ficial when the number of noisy kernels is increased).

In summary, this experiment should remind us that semantic classification setups use a substantial amount of prior knowledge. Prior knowledge implies a *pre-selection* of highly effective kernels—a carefully chosen set of strong kernels constitutes a bias towards the sum kernel. Clearly, pre-selection of strong kernels reduces the need for learning kernel weights; however, in settings where prior knowledge is sparse, statistical (or even adaptive, adversarial) noise is inherently contained in the feature extraction—thus, beneficial effects of MKL are expected to be more pronounced in such a scenario.

### 4.3 One Argument for Learning the Multiple Kernel Weights: Varying Informative Subsets of Data

In the previous sections, we presented evidence for why the sum-kernel SVM is considered to be a strong learner in visual image categorization. Nevertheless, in our experiments we observed gains in accuracy by using MKL

for many concepts. In this section, we investigate causes for this performance gain.

Intuitively speaking, one can claim that the kernel non-uniformly contain varying amounts of information content. We investigate more specifically what information content this is and why it differs over the kernels. Our main hypothesis is that common kernels in visual concept classification are informative with respect to varying subsets of the data. This stems from features being frequently computed from many combinations of color channels. We can imagine that blue color present in the upper third of an image can be crucial for prediction of photos having clear sky, while other photos showing a sundown or a smoggy sky tend to contain white or yellow colors; this means that a particular kernel / feature group can be crucial for some images, while it may be almost useless—or even counterproductive—for others.

However, the information content is accessed by MKL via the quantity given by Eq. (8); the latter is a *global* information measure, which is computed over the support vectors (which in turn are chosen over the *whole dataset*). In other words, the kernel weights are global weights that uniformly hold in all regions of the input space. Explicitly finding informative subsets of the input space on real data may not only imply a too high computational burden (note that the number of partitions of an  $n$ -element training set is exponentially in  $n$ ) but also is very likely to lead to overfitting.

To understand the implications of the above to computer vision, we performed the following toy experiment. We generated a fraction of  $p_+ = 0.25$  of positively labeled and  $p_- = 0.75$  of negatively labeled  $6m$ -dimensional training examples (motivated by the unbalancedness of training sets usually encountered in computer vision) in the following way: the features were divided in  $k$  feature groups each consisting of six features. For each feature group, we split the training set into an informative and an uninformative set (the size is varying over the feature groups); thereby, the informative sets of the particular feature groups are disjoint. Subsequently, each feature group is processed by a Gaussian kernel, where the width is determined heuristically in the same way as in the real experiments shown earlier in this paper.

Thereby, we consider two experimental setups for sampling the data, which differ in the number of employed kernels  $m$  and the sizes of the informative sets. In both

cases, the informative features are drawn from two sufficiently distant normal distributions (one for each class) while the uninformative features are just Gaussian noise (mixture of Gaussians). The experimental setup of the first experiment can be summarized as follows:

**Experimental Settings for Experiment 1 (3 kernels):**

$$n_{k=1,2,3} = (300, 300, 500), p_+ := P(y = +1) = 0.25 \quad (9)$$

The features for the informative subset are drawn according to

$$f_i^{(k)} \sim \begin{cases} N(0.0, \sigma_k) & \text{if } y_i = -1 \\ N(0.4, \sigma_k) & \text{if } y_i = +1 \end{cases} \quad (10)$$

$$\sigma_k = \begin{cases} 0.3 & \text{if } k = 1, 2 \\ 0.4 & \text{if } k = 3 \end{cases} \quad (11)$$

The features for the uninformative subset are drawn according to

$$f^{(k)} \sim (1 - p_+)N(0.0, 0.5) + p_+N(0.4, 0.5). \quad (12)$$

For Experiment 1 the three kernels had disjoint informative subsets of sizes  $n_{k=1,2,3} = (300, 300, 500)$ . We used 1100 data points for training and the same amount for testing. We repeated this experiment 500 times with different random draws of the data.

Note that the features used for the uninformative subsets are drawn as a mixture of the Gaussians used for the informative subset, but with a higher variance, though. The increased variance encodes the assumption that the feature extraction produces unreliable results on the uninformative data subset. None of these kernels are pure noise or irrelevant. Each kernel is the best one for its own informative subset of data points.

We now turn to the experimental setup of the second experiment:

**Experimental Settings for Experiment 2 (5 kernels):**

$$n_{k=1,2,3,4,5} = (300, 300, 500, 200, 500), \\ p_+ := P(y = +1) = 0.25$$

The features for the informative subset are drawn according to

$$f_i^{(k)} \sim \begin{cases} N(0.0, \sigma_k) & \text{if } y_i = -1 \\ N(m_k, \sigma_k) & \text{if } y_i = +1 \end{cases} \quad (13)$$

Experiment	$\ell_\infty$ -SVM	$\ell_{1.0625}$ -MKL	t-test p-value
1	$68.72 \pm 3.27$	$69.49 \pm 3.17$	0.000266
2	$55.07 \pm 2.86$	$56.39 \pm 2.84$	$4.7 \cdot 10^{-6}$

Table 8: Varying Informative Subsets of Data: AP Scores in Toy experiment using Kernels with disjoint informative subsets of Data

$$m_k = \begin{cases} 0.4 & \text{if } k = 1, 2, 3 \\ 0.2 & \text{if } k = 4, 5 \end{cases} \quad (14)$$

$$\sigma_k = \begin{cases} 0.3 & \text{if } k = 1, 2 \\ 0.4 & \text{if } k = 3, 4, 5 \end{cases} \quad (15)$$

The features for the uninformative subset are drawn according to

$$f^{(k)} \sim (1 - p_+)N(0.0, 0.5) + p_+N(m_k, 0.5) \quad (16)$$

As for the real experiments, we normalized the kernels to having standard deviation 1 in Hilbert space and optimized the regularization constant by grid search in  $C \in \{10^i \mid i = -2, -1.5, \dots, 2\}$ .

Table 8 shows the results. The null hypothesis of equal means is rejected by a t-test with a p-value of 0.000266 and 0.0000047, respectively, for Experiment 1 and 2, which is highly significant.

The design of the Experiment 1 is no exceptional lucky case: we observed similar results when using more kernels; the performance gaps then even increased. Experiment 2 is a more complex version of Experiment 1 using five kernels instead of just three. Again, the informative subsets are disjoint, but this time of sizes 300, 300, 500, 200, and 500; the the Gaussians are centered at 0.4, 0.4, 0.4, 0.2, and 0.2, respectively, for the positive class; and the variance is taken as  $\sigma_k = (0.3, 0.3, 0.4, 0.4, 0.4)$ . Compared to Experiment 1, this results in even bigger performance gaps between the sum-kernel SVM and the non-sparse  $\ell_{1.0625}$ -MKL. One can imagine to create learning scenarios with more and more kernels in the above way, thus increasing the performance gaps—since we aim at a relative comparison, this, however, would not further contribute to validating or rejecting our hypothesis.

Furthermore, we also investigate the single-kernel performance of each kernel: we observed the best single-kernel SVM (which attained AP scores of 43.60, 43.40,

and 58.90 for Experiment 1) being inferior to both MKL (regardless of the employed norm parameter  $p$ ) and the sum-kernel SVM. The differences were significant with fairly small p-values (for example, for  $\ell_{1.25}$ -MKL the p-value was about 0.02).

We emphasize that we did not design the example in order to achieve a maximal performance gap between the non sparse MKL and its competitors. For such an example, see the toy experiment of [21], which is replicated in the supplemental material including additional analysis. Our focus here was to confirm our hypothesis that kernels in semantic concept classification are based on varying subsets of the data—although MKL computes global weights, it emphasizes on kernels that are relevant on the largest informative set and thus approximates the infeasible combinatorial problem of computing an optimal partition/grid of the space into regions which underlie identical optimal weights. Though, in practice, we expect the situation to be more complicated as informative subsets may overlap between kernels.

Nevertheless, our hypothesis also opens the way to new directions for learning of kernel weights, namely restricted to subsets of data chosen according to a meaningful principle. Finding such principles is one the future goals of MKL—we sketched one possibility: locality in feature space. A first starting point may be the work of [51, 52] on localized MKL.

## 5 Conclusions

When measuring data with different measuring devices, it is always a challenge to combine the respective devices’ uncertainties in order to fuse all available sensor information optimally. In this paper, we revisited this important topic and discussed machine learning approaches to adaptively combine different image descriptors in a systematic and theoretically well founded manner. While MKL approaches in principle solve this problem it has been observed that the standard  $\ell_1$ -norm based MKL often cannot outperform SVMs that use an average of a large number of kernels. One hypothesis why this seemingly unintuitive result may occur is that the sparsity prior may not be appropriate in many real world problems—especially, when prior knowledge is already at hand. We tested whether this hypothesis holds true for computer vision and applied the

recently developed non-sparse  $\ell_p$  MKL algorithms to object classification tasks. The  $\ell_p$ -norm constitutes a slightly less severe method of sparsification. By choosing  $p$  as a hyperparameter, which controls the degree of non-sparsity and regularization, from a set of candidate values with the help of a validation data, we showed that  $\ell_p$ -MKL significantly improves SVMs with averaged kernels and the standard sparse  $\ell_1$  MKL.

Future work will study localized MKL and methods to include hierarchically structured information into MKL, e.g. knowledge from taxonomies, semantic information or spatial priors. Another interesting direction is MKL-KDA [27, 28]. The difference to the method studied in the present paper lies in the base optimization criterion: KDA [53] leads to non-sparse solutions in  $\alpha$  while ours leads to sparse ones (i.e., a low number of support vectors). While on the computational side the latter is expected to be advantageous, the first one might lead to more accurate solutions. We expect the regularization over kernel weights (i.e., the choice of the norm parameter  $p$ ) having similar effects for MKL-KDA like for MKL-SVM. Future studies will expand on that topic.<sup>1</sup>

## Acknowledgments

This work was supported in part by the Federal Ministry of Economics and Technology of Germany (BMWi) under the project THESEUS (FKZ 01MQ07018), by Federal Ministry of Education and Research (BMBF) under the project REMIND (FKZ 01-IS07007A), by the Deutsche Forschungsgemeinschaft (DFG), and by the FP7-ICT program of the European Community, under the PASCAL2 Network of Excellence (ICT-216886). Marius Kloft acknowledges a scholarship by the German Academic Exchange Service (DAAD).

## References

- [1] Vapnik V (1995) The Nature of Statistical Learning Theory. New York: Springer.

<sup>1</sup>First experiments on ImageCLEF2010 show for sum kernel SRKDA [54] a result of 39.29 AP points which is slightly better than the sum kernel results for the SVM (39.11 AP) but worse than MKL-SVM.



- [2] Cortes C, Vapnik V (1995) Support-vector networks. In: Machine Learning. pp. 273–297.
- [3] Vapnik VN (1998) Statistical Learning Theory. Wiley-Interscience.
- [4] Chapelle O, Haffner P, Vapnik V (1999) SVMs for histogram-based image classification. IEEE Trans on Neural Networks 10: 1055–1064.
- [5] Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B (2001) An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks 12: 181–201.
- [6] Schölkopf B, Smola AJ (2002) Learning with Kernels. Cambridge, MA: MIT Press.
- [7] Jaakkola T, Haussler D (1998) Exploiting generative models in discriminative classifiers. In: Advances in Neural Information Processing Systems. volume 11, pp. 487–493.
- [8] Zien A, Rätsch G, Mika S, Schölkopf B, Lengauer T, et al. (2000) Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics 16: 799–807.
- [9] Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, volume 2, pp. 2169–2178.
- [10] Zhang J, Marszalek M, SLazebnik, Schmid C (2007) Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73: 213–238.
- [11] Kumar A, Sminchisescu C (2007) Support kernel machines for object recognition. In: IEEE International Conference on Computer Vision.
- [12] Gehler PV, Nowozin S (2009) On feature combination for multiclass object classification. In: ICCV. IEEE, pp. 221–228.
- [13] Lanckriet GR, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI (2004) Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research : 27–72.
- [14] Bach F, Lanckriet G, Jordan M (2004) Multiple kernel learning, conic duality, and the smo algorithm. International Conference on Machine Learning .
- [15] Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B (2006) Large Scale Multiple Kernel Learning. Journal of Machine Learning Research 7: 1531–1565.
- [16] Rakotomamonjy A, Bach F, Canu S, Grandvalet Y (2008) SimpleMKL. Journal of Machine Learning Research 9: 2491–2521.
- [17] Cortes C, Gretton A, Lanckriet G, Mohri M, Rostamizadeh A (2008). Proceedings of the NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels. URL [http://www.cs.nyu.edu/learning/\\_kernels](http://www.cs.nyu.edu/learning/_kernels).
- [18] Kloft M, Brefeld U, Laskov P, Sonnenburg S (2008) Non-sparse multiple kernel learning. In: Proc. of the NIPS Workshop on Kernel Learning: Automatic Selection of Kernels.
- [19] Cortes C, Mohri M, Rostamizadeh A (2009) L2 regularization for learning kernels. In: Proceedings of the International Conference on Uncertainty in Artificial Intelligence.
- [20] Kloft M, Brefeld U, Sonnenburg S, Laskov P, Müller KR, et al. (2009) Efficient and accurate lp-norm multiple kernel learning. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A, editors, Advances in Neural Information Processing Systems 22, MIT Press. pp. 997–1005.
- [21] Kloft M, Brefeld U, Sonnenburg S, Zien A (2011) Lp-norm multiple kernel learning. Journal of Machine Learning Research 12: 953–997.
- [22] Orabona F, Luo J, Caputo B (2010) Online-batch strongly convex multi kernel learning. In: CVPR. pp. 787–794.
- [23] Vedaldi A, Gulshan V, Varma M, Zisserman A (2009) Multiple kernels for object detection. In: Computer Vision, 2009 IEEE 12th International Conference on. pp. 606–613. doi:10.1109/ICCV.2009.5459183.
- [24] Galleguillos C, McFee B, Belongie SJ, Lanckriet GRG (2010) Multi-class object localization by combining local contextual interactions. In: CVPR. pp. 113–120.
- [25] Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2009). The PASCAL Visual Object Classes Challenge 2009 (VOC2009). <http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html>.
- [26] Nowak S, Huiskes MJ (2010) New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. In: CLEF (Notebook Papers/LABs/Workshops).
- [27] Yan F, Kittler J, Mikolajczyk K, Tahir A (2009) Non-sparse multiple kernel learning for fisher discriminant analysis. In: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining. Washington, DC, USA:

- IEEE Computer Society, ICDM '09, pp. 1064–1069. doi: 10.1109/ICDM.2009.84.
- [28] Yan F, Mikolajczyk K, Barnard M, Cai H, Kittler J (2010) Lp norm multiple kernel fisher discriminant analysis for object and image categorisation. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 0: 3626–3632.
  - [29] Cao L, Luo J, Liang F, Huang TS (2009) Heterogeneous feature machines for visual recognition. In: *ICCV*. pp. 1095–1102.
  - [30] Gehler PV, Nowozin S (2009) Let the kernel figure it out; principled learning of pre-processing for kernel classifiers. In: *CVPR*. pp. 2836–2843.
  - [31] Varma M, Babu BR (2009) More generality in efficient multiple kernel learning. In: *ICML*. p. 134.
  - [32] Zien A, Ong C (2007) Multiclass multiple kernel learning. In: *ICML*. pp. 1191–1198.
  - [33] Rakotomamonjy A, Bach F, Canu S, Grandvalet Y (2007) More efficiency in multiple kernel learning. In: *ICML*. pp. 775–782.
  - [34] Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J (2002) On kernel-target alignment. In: *Advances in Neural Information Processing Systems*. volume 14, pp. 367–373.
  - [35] Cortes C, Mohri M, Rostamizadeh A (2010) Two-stage learning kernel algorithms. In: Fürnkranz J, Joachims T, editors, *ICML*. Omnipress, pp. 239–246.
  - [36] Mika S, Rätsch G, Weston J, Schölkopf B, Smola A, et al. (2003) Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 25: 623 – 628.
  - [37] Marszalek M, Schmid C. Learning representations for visual object class recognition. URL <http://pascallin.ecs.soton.ac.uk/\\challenges/VOC/voc2007/workshop/\\marszalek.pdf>.
  - [38] Lampert C, Blaschko M (2008) A multiple kernel learning approach to joint multi-class object detection. In: *DAGM*. pp. 31–40.
  - [39] Csurka G, Bray C, Dance C, Fan L (2004) Visual categorization with bags of keypoints. In: *Workshop on Statistical Learning in Computer Vision, ECCV*. Prague, Czech Republic, pp. 1–22.
  - [40] Lowe D (2004) Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision* 60: 91–110.
  - [41] van de Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. *IEEE Trans Pat Anal & Mach Intel* .
  - [42] Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Diego, USA, volume 1, pp. 886–893.
  - [43] Canny J (1986) A computational approach to edge detection. *IEEE Trans on Pattern Analysis and Machine Intelligence* 8: 679–714.
  - [44] Zien A, Ong CS (2007) Multiclass multiple kernel learning. In: *Proceedings of the 24th international conference on Machine learning (ICML)*. ACM, pp. 1191–1198.
  - [45] Chapelle O, Rakotomamonjy A (2008) Second order optimization of kernel parameters. In: *Proc. of the NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*.
  - [46] Sonnenburg S, Rätsch G, Henschel S, Widmer C, Behr J, et al. (2010) The shogun machine learning toolbox. *Journal of Machine Learning Research* .
  - [47] Kloft M, Brefeld U, Sonnenburg S, Zien A (2010) Non-sparse regularization for multiple kernel learning. *Journal of Machine Learning Research* .
  - [48] Moosmann F, Nowak E, Jurie F (2008) Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 30: 1632–1646.
  - [49] Moosmann F, Triggs B, Jurie F (2006) Fast discriminative visual codebooks using randomized clustering forests. In: *Advances in Neural Information Processing Systems*.
  - [50] Breiman L (1996) Bagging predictors. *Mach Learn* 24: 123–140.
  - [51] Gönen M, Alpaydm E (2010) Localized multiple kernel regression. In: *Proceedings of the 20th IAPR International Conference on Pattern Recognition*.
  - [52] Yang J, Li Y, Tian Y, Duan L, Gao W (2009) Group-sensitive multiple kernel learning for object categorization. In: *ICCV*. pp. 436–443.
  - [53] Mika S, Rätsch G, Weston J, Schölkopf B, Müller KR (1999) Fisher discriminant analysis with kernels. In: Hu YH, Larsen J, Wilson E, Douglas S, editors, *Neural Networks for Signal Processing IX*. IEEE, pp. 41–48.
  - [54] Cai D, He X, Han J (2007) Efficient kernel discriminant analysis via spectral regression. In: *Proc. Int. Conf. on Data Mining (ICDM'07)*.